

Adaptive Survival Trials

Dominic Magirr¹, Thomas Jaki², Franz Koenig¹, and Martin Posch¹

¹Section for Medical Statistics, Medical University of Vienna, Austria

²Department of Mathematics and Statistics, Lancaster University, UK

May 8, 2014

Abstract

Mid-study design modifications are becoming increasingly accepted in confirmatory clinical trials, so long as appropriate methods are applied such that error rates are controlled. It is therefore unfortunate that the important case of time-to-event endpoints is not easily handled by the standard theory. We analyze current methods that allow design modifications to be based on the full interim data, i.e., not only the observed event times but also secondary endpoint and safety data from patients who are yet to have an event. We show that the final test statistic may ignore a substantial subset of the observed event times. Since it is the data corresponding to the earliest recruited patients that is ignored, this neglect becomes egregious when there is specific interest in learning about long-term survival. An alternative test incorporating all event times is proposed, where a conservative assumption is made in order to guarantee type I error control. We examine the properties of our proposed approach using the example of a clinical trial comparing two cancer therapies.

Keywords: Adaptive design; Brownian motion; Clinical trial; Combination test; Sample size reassessment; Time-to-event.

1 Introduction

There are often strong ethical and economic arguments for conducting interim analyses of an ongoing clinical trial and for making changes to the design if warranted by the accumulating data. One may decide, for example, to increase the sample size on the basis of promising interim results. Or perhaps one might wish to drop a treatment from a multi-arm study on the basis of unsatisfactory safety data. Owing to the complexity of clinical drug development, it is not always possible to anticipate the need for such modifications, and therefore not all contingencies can be dealt with in the statistical design.

Unforeseen interim modifications complicate the (frequentist) statistical analysis of the trial considerably. Over recent decades many authors have investigated so-called “adaptive designs” in an effort to maintain the concept of type I error control (Bauer and Köhne, 1994; Proschan and Hunsberger, 1995; Müller and Schäfer, 2001; Hommel, 2001). Although Bayesian adaptive methods are becoming increasingly popular, type I error control is still deemed important in the setting of a confirmatory phase III trial (Berry et al., 2010, p. 6), and recent years have seen hybrid adaptive designs proposed, whereby the interim decision is based on Bayesian methods, but the final hypothesis test remains frequentist (Brannath et al., 2009; Di Scala and Glimm, 2011).

While the theory of adaptive designs is now well understood if responses are observed immediately, subtle problems arise when responses are delayed, e.g., in survival trials.

Schäfer and Müller (2001) proposed adaptive survival tests that are constructed using the independent increments property of logrank test statistics (c.f., Wassmer, 2006; Desseaux and Porcher, 2007; Jahn-Eimermacher and Ingel, 2009). However, as pointed out by Bauer and Posch (2004), these methods only work if interim decision making is based solely on the interim logrank test statistics and any secondary endpoint data from patients who have already had an event. In other words, investigators must remain blind to the data from patients who are censored at the interim analysis. Irle and Schäfer (2012) argue that decisions regarding interim design modifications should be as substantiated as possible, and propose a test procedure that allows investigators to use the full interim data. This methodology, similar to that of Jenkins et al. (2011), does not require any assumptions regarding the joint distribution of survival times and short-term secondary endpoints, as do, e.g., the methods proposed by Stallard (2010), Friede et al. (2011, 2012) and Hampson and Jennison (2013).

The first goal of this article is to clarify the proposals of Jenkins et al. (2011) and Irle and Schäfer (2012), showing that they are both based on weighted inverse-normal test statistics (Lehmacher and Wassmer, 1999), with the common disadvantage that the final test statistic may ignore a substantial subset of the observed survival times. This is a serious limitation, as disregarding part of the observed data is generally considered inappropriate even if statistical error probabilities are controlled – see, for example, the discussion on overrunning in group sequential trials (Hampson and Jennison, 2013). Our secondary goal is therefore to propose an alternative test that retains the strict type I error control and flexibility of the aforementioned designs, but bases the final test decision on a statistic that takes into account all available survival times. As ever, there is no free lunch, and the assumption that we require to ensure type I error control induces a certain amount of conservatism. We evaluate the properties of our proposed approach using the example of a clinical trial comparing two cancer therapies.

2 Adaptive Designs

2.1 Standard theory

A comprehensive account of adaptive design methodology can be found in Bretz et al. (2009). For testing a null hypothesis, $H_0 : \theta = 0$, against the one-sided alternative, $H_a : \theta > 0$, the archetypal two-stage adaptive test statistic is of the form $f_1(p_1) + f_2(p_2)$, where p_1 is the p-value based on the first-stage data, p_2 is the p-value from the (possibly adapted) second-stage test, and f_1 and f_2 are prespecified monotonically decreasing functions. Consider the simplest case that no early rejection of the null hypothesis is possible at the end of the first stage. The null hypothesis is rejected at level α whenever $f_1(p_1) + f_2(p_2) > k$, where k satisfies

$$\int_0^1 \int_0^1 \mathbf{1}\{f_1(p_1) + f_2(p_2) \leq k\} dp_1 dp_2 = 1 - \alpha.$$

In their seminal paper, Bauer and Köhne (1994) took $f_i(p_i) = -\log(p_i)$ for $i = 1, 2$. We will restrict attention to the weighted inverse-normal test statistic (Lehmacher and Wassmer, 1999),

$$Z = w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2), \quad (1)$$

where Φ denotes the standard normal distribution function and w_1 and w_2 are prespecified weights such that $w_1^2 + w_2^2 = 1$. If $Z > \Phi^{-1}(1 - \alpha)$, then H_0 may be rejected at level α . The assumptions required to make this a valid level- α test are as follows (see Brannath et al., 2012).

Assumption 1

Let X_1^{int} denote the data available at the interim analysis, where $X_1^{\text{int}} \in \mathbb{R}^n$ with distribution function $G(x_1^{\text{int}}; \theta)$. The calendar time of the interim analysis will be denoted T^{int} . In general, X_1^{int} will contain information not only concerning the primary endpoint, but also measurements on secondary endpoints and safety data. It is assumed that the first-stage p-value function $p_1 : \mathbb{R}^n \rightarrow [0, 1]$ satisfies

$$\int_{\mathbb{R}^n} \mathbf{1}\{p_1(x_1^{\text{int}}) \leq u\} dG(x_1^{\text{int}}; 0) \leq u \text{ for all } u \in [0, 1].$$

Assumption 2

At the interim analysis, a second-stage design d is chosen. The second-stage design is allowed to depend on the unblinded first-stage data without prespecifying an adaptation rule. Denote the second-stage data by Y , where $Y \in \mathbb{R}^m$. It is assumed that the distribution function of Y , denoted by $F_{d, x_1^{\text{int}}}(y, \theta)$, is known for all possible second stage designs, d , and all first-stage outcomes, x_1^{int} .

Assumption 3

The second-stage p-value function $p_2 : \mathbb{R}^m \rightarrow [0, 1]$ satisfies $\int_{\mathbb{R}^m} \mathbf{1}\{p_2(y) \leq u\} dF_{d, x_1^{\text{int}}}(y; 0) \leq u$ for all $u \in [0, 1]$.

2.2 Immediate responses

The aforementioned assumptions are easy to justify when primary endpoint responses are observed more-or-less immediately. In this case X_1^{int} contains the responses of all patients recruited prior to the interim analysis. A second-stage design d can subsequently be chosen with the responses from a new cohort of patients contributing to Y (Figure 1).

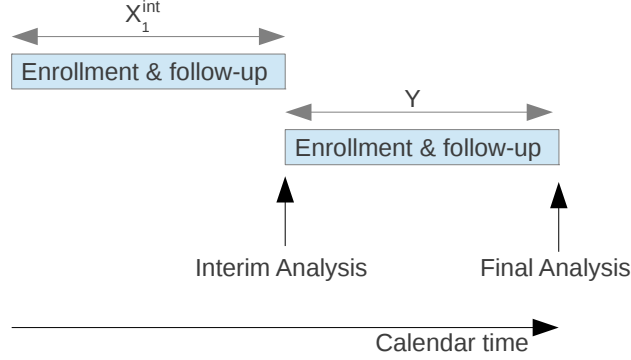


Figure 1: Schematic of a standard two-stage adaptive trial with immediate response.

2.3 Delayed responses and the independent increments assumption

An interim analysis may take place whilst some patients have entered the study but have yet to provide a data point on the primary outcome measure. Most approaches to this problem (e.g., Schäfer and Müller, 2001; Wassmer, 2006; Jahn-Eimermacher and Ingel, 2009) attempt to take advantage of the well known independent increments structure of score statistics in group sequential designs (Jennison and Turnbull, 2000). As pictured in Figure 2, X_1^{int} will generally include responses on short-term secondary endpoints and safety data from patients who are yet to provide a primary outcome measure, while Y consists of some delayed responses from patients recruited prior to T^{int} , mixed together with responses from a new cohort of patients.

Let $S(X_1^{\text{int}})$ and $\mathcal{I}(X_1^{\text{int}})$ denote the score statistic and Fisher's information for θ , calculated from primary endpoint responses in X_1^{int} . Assuming suitable regularity conditions, the asymptotic null distribution of $S(X_1^{\text{int}})$ is Gaussian with mean zero and variance $\mathcal{I}(X_1^{\text{int}})$ (Cox and Hinkley, 1979, p. 107). The independent increments assumption is that for all first-stage outcomes x_1^{int} and second-stage designs d , the null distribution of Y is such that

$$S(x_1^{\text{int}}, Y) - S(x_1^{\text{int}}) \sim \mathcal{N}\{0, \mathcal{I}(x_1^{\text{int}}, Y) - \mathcal{I}(x_1^{\text{int}})\}, \quad (2)$$

at least approximately, where $S_{X_1^{\text{int}}, Y}$ and $\mathcal{I}_{X_1^{\text{int}}, Y}$ denote the score statistic and Fisher's information for θ , calculated from primary endpoint responses in (X_1^{int}, Y) .

Unfortunately, (2) is seldom realistic in an adaptive setting. Bauer and Posch (2004) show that if the adaptive strategy at the interim analysis is dependent on short-term outcomes in X_1^{int} that are correlated with primary endpoint outcomes in Y , i.e., from the same patient, then a naive appeal to the independent increments assumption can lead to very large type I error inflation.

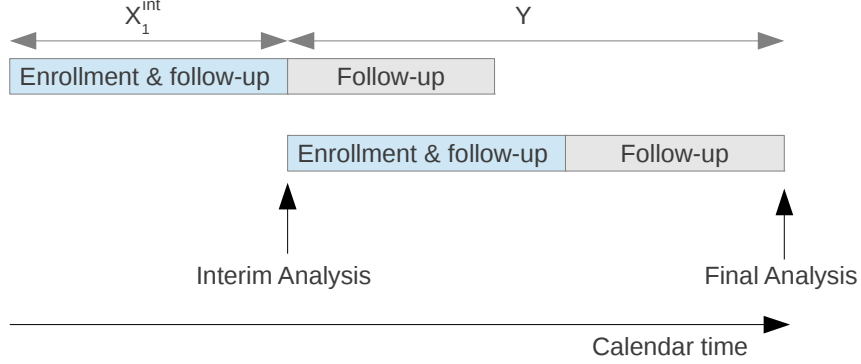


Figure 2: Schematic of a two-stage adaptive trial with delayed response under the independent increments assumption.

2.4 Delayed responses with “patient-wise separation”

An alternative approach, which we shall coin “patient-wise separation”, redefines the first-stage p-value, $p_1 : \mathbb{R}^p \rightarrow [0, 1]$, to be a function of X_1 , where X_1 denotes all the data from patients recruited prior to T^{int} , followed-up until calendar time T^{max} – which corresponds to the prefixed maximum duration of the trial. It is assumed that X_1 takes values in \mathbb{R}^p according to distribution function $\tilde{G}(x_1; \theta)$. Assumption 1 is replaced with:

$$\int_{\mathbb{R}^p} \mathbf{1}\{p_1(x_1) \leq u\} d\tilde{G}(x_1; 0) \leq u \text{ for all } u \in [0, 1]. \quad (3)$$

In this case p_1 may not be observable at the time the second-stage design d is chosen. This is not a problem, as long as no early rejection at the end of the first stage is foreseen. Any interim decisions, such as increasing the sample size, do not require any knowledge of p_1 . It is assumed that Y consists of responses from a new cohort of patients, such that x_1^{int} could be formally replaced with x_1 in assumptions 2 and 3. We call this “patient-wise separation” because data from the same patient cannot contribute to both p_1 and p_2 .

Liu and Pledger (2005) consider such an approach for a clinical trial where a patient’s primary outcome is measured after a fixed period of follow-up, e.g., 4 months. Provided that one is willing to wait for all responses, it is straightforward to prespecify a first-stage p-value function such that (3) holds.

For an adaptive trial with a time-to-event endpoint, however, one must be very careful to ensure that (3) holds, as one is typically not prepared to wait for all first-stage patients – those patients recruited prior to T^{int} – to have an event. Rather, p_1 is defined as the p-value from an, e.g., logrank test applied to the data from first-stage patients followed up until time T_1 , for some $T_1 < T^{\text{max}}$. In this case it is vital that T_1 be fixed at the start of the trial, either explicitly or implicitly (Jenkins

et al., 2011; Irle and Schäfer, 2012). Otherwise, if T_1 were to depend on the adaptive strategy at the interim analysis, this would impact the distribution of p_1 and could lead to type I error inflation.

The situation is represented pictorially in Figure 3. An unfortunate consequence of prefixing T_1 is that this will not, in all likelihood, correspond to the end of follow-up for second-stage patients. All events of first-stage patients that occur after T_1 make no contribution to the statistic (1); they are “thrown away”.

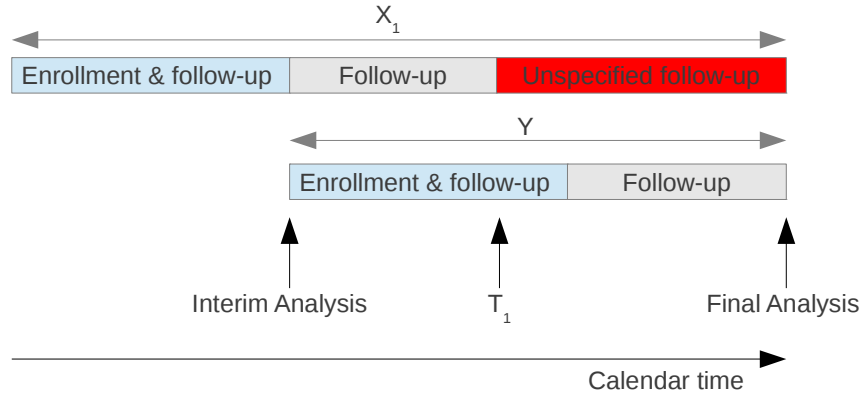


Figure 3: Schematic of a two-stage adaptive trial with “patient-wise separation”.

3 Adaptive Survival Studies

3.1 Jenkins et al. (2011) method

Consider a randomized clinical trial comparing survival times on an experimental treatment, E , with those on a control treatment, C . We will focus on the logrank statistic for testing the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $H_a : \theta > 0$, where θ is the log hazard ratio, assuming proportional hazards. Let $D_1(t)$ and $S_1(t)$ denote the number of uncensored events and the usual logrank score statistic, respectively, based on the data from first-stage patients – those patients recruited prior to the interim analysis – followed up until calendar time t , $t \in [0, T^{\max}]$. Under the null hypothesis, assuming equal allocation and a large number of events, the variance of $S_1(t)$ is approximately equal to $D_1(t)/4$ (e.g., Whitehead, 1997, Section 3.4). The first-stage p-value must be calculated at a prefixed time point T_1 :

$$p_1 = 1 - \Phi \left[2 \{S_1(T_1)\} / \{D_1(T_1)\}^{1/2} \right]. \quad (4)$$

There are two possible ways of specifying T_1 in (4). From a practical perspective, a *calendar time* approach is often attractive as T_1 is specified explicitly, which facilitates straightforward planning.

On the other hand, this can produce a misspowered study if the recruitment rate and/or survival times differ markedly from those anticipated. An *event driven* approach may be preferred, whereby the number of events is prefixed at d_1 , say, and

$$T_1 := \min \{t : D_1(t) = d_1\}. \quad (5)$$

Jenkins et al. (2011) describe a “patient-wise separation” adaptive survival trial, with test statistic (1), first-stage p-value (4) and T_1 defined as in (5). While their focus is on subgroup selection, we will appropriate their method for the simpler situation of a single comparison, where at the interim analysis one has the possibility to alter the pre-planned number of events from second-stage patients – i.e., those patients recruited post T^{int} . All that remains to be specified at the design stage is the choice of weights w_1 and w_2 . It is anticipated that p_2 will be the p-value corresponding to a logrank test based on second-stage patients, i.e.,

$$p_2 = 1 - \Phi \left[2S_2(T_2^*) / \{D_2(T_2^*)\}^{1/2} \right],$$

where $T_2^* := \min \{t : D_2(t) = d_2^*\}$ with $S_2(t)$ and $D_2(t)$ defined analogously to $S_1(t)$ and $D_1(t)$, and d_2^* is to be specified at the interim analysis. Ideally, the weights should be chosen in proportion to the information (number of events) contributed from each stage. In an adaptive trial, it is impossible to achieve the correct weighting in every scenario. Jenkins et al. prespecify the *envisioned* number of second-stage events, d_2 , and choose weights $w_1 = \{d_1/(d_1 + d_2)\}^{1/2}$ and $w_2 = \{d_2/(d_1 + d_2)\}^{1/2}$.

3.2 Irle and Schäfer (2012) method

Irle and Schäfer (2012) propose an alternative procedure. Instead of explicitly combining stage-wise p-values, they employ the closely related *conditional error* approach (Proschan and Hunsberger, 1995; Posch and Bauer, 1999; Müller and Schäfer, 2001).

They begin by prespecifying a level- α test with decision function, φ , taking values in $\{0, 1\}$ corresponding to nonrejection and rejection of H_0 , respectively. For a survival trial, this entails specifying the sample size, duration of follow-up, test statistic, recruitment rate, etc. Then, at some (not necessarily prespecified) timepoint, T^{int} , an interim analysis is performed. The timing of the interim analysis induces a partition of the trial data, (X_1, X_2) , where X_1 and X_2 denote the data from patients recruited prior- T^{int} and post- T^{int} , respectively, followed-up until time T^{max} . More specifically, Irle and Schäfer (2012) suggest the decision function

$$\varphi(X_1, X_2) = \mathbf{1} \left[2S_{1,2}(T_{1,2}) / \{D_{1,2}(T_{1,2})\}^{1/2} > \Phi^{-1}(1 - \alpha) \right], \quad (6)$$

where $D_{1,2}(T_{1,2})$ and $S_{1,2}(T_{1,2})$ denote the number of uncensored events and the usual logrank score statistic, respectively, based on data from all patients (from both stages) followed-up until time $T_{1,2}$, where $T_{1,2} := \min \{t : D_{1,2}(t) = d_{1,2}\}$ for some prespecified number of events $d_{1,2}$.

At the interim analysis, the general idea is to use the unblinded first-stage data x_1^{int} to define a second-stage design, d , without the need for a prespecified adaptation strategy. Again, the definition of d includes factors such as sample size, follow-up period, recruitment rate, etc., in addition to a second-stage decision function $\psi_{x_1^{\text{int}}} : \mathbb{R}^m \rightarrow \{0, 1\}$ based on second-stage data

$Y \in \mathbb{R}^m$. Irle and Schäfer (2012) focus their attention on a specific design change; namely, the possibility of increasing the number of events from $d_{1,2}$ to $d_{1,2}^*$ by extending the follow-up period. They assume that $Y := (X_1, X_2) \setminus X_1^{\text{int}}$ and propose the second-stage decision function

$$\psi_{x_1^{\text{int}}}(Y) = \mathbf{1} \left[2S_{1,2}(T_{1,2}^*) / \{D_{1,2}(T_{1,2}^*)\}^{1/2} \geq b^* \right], \quad (7)$$

where $T_{1,2}^* := \min \{t : D_{1,2}(t) = d_{1,2}^*\}$ and b^* is a cutoff value that must be determined. Ideally, one would like to choose b^* such that $E_{H_0}(\psi_{X_1^{\text{int}}} | X_1^{\text{int}} = x_1^{\text{int}}) = E_{H_0}(\varphi | X_1^{\text{int}} = x_1^{\text{int}})$, as this would ensure that

$$E_{H_0}(\psi_{X_1^{\text{int}}}) = E_{H_0} \left\{ E_{H_0} \left(\psi_{X_1^{\text{int}}} | X_1^{\text{int}} \right) \right\} = E_{H_0} \left\{ E_{H_0} (\varphi | X_1^{\text{int}}) \right\} = E_{H_0}(\varphi) = \alpha, \quad (8)$$

i.e., the overall procedure controls the type I error rate at level α . Unfortunately, this approach is not directly applicable in a survival trial where X_1^{int} contains short-term data from first-stage patients surviving beyond T^{int} . This is because it is impossible to calculate $E_{H_0}(\varphi | X_1^{\text{int}} = x_1^{\text{int}})$ and $E_{H_0}(\psi_{X_1^{\text{int}}} | X_1^{\text{int}} = x_1^{\text{int}})$, owing to the unknown joint distribution of survival times and the secondary/safety endpoints already observed at the interim analysis, c.f. Section 2.3. Irle and Schäfer (2012) get around this problem by conditioning on additional variables; namely, $S_1(T_{1,2})$ and $S_1(T_{1,2}^*)$. Choosing $\psi_{x_1^{\text{int}}}$ such that

$$E_{H_0} \left\{ \psi_{X_1^{\text{int}}} | X_1^{\text{int}} = x_1^{\text{int}}, S_1(T_{1,2}) = s_1, S_1(T_{1,2}^*) = s_1^* \right\} = E_{H_0} \left\{ \varphi | X_1^{\text{int}} = x_1^{\text{int}}, S_1(T_{1,2}) = s_1, S_1(T_{1,2}^*) = s_1^* \right\}$$

ensures that $E_{H_0}(\psi_{X_1^{\text{int}}}) = \alpha$ following the same argument as (8).

Irle and Schäfer (2012) show that, asymptotically,

$$E_{H_0} \left\{ \varphi | X_1^{\text{int}} = x_1^{\text{int}}, S_1(T_{1,2}) = s_1, S_1(T_{1,2}^*) = s_1^* \right\} = E_{H_0} \left\{ \varphi | S_1(T_{1,2}) = s_1 \right\}$$

and

$$E_{H_0} \left\{ \psi_{X_1^{\text{int}}} | X_1^{\text{int}} = x_1^{\text{int}}, S_1(T_{1,2}) = s_1, S_1(T_{1,2}^*) = s_1^* \right\} = E_{H_0} \left\{ \psi_{x_1^{\text{int}}} | S_1(T_{1,2}^*) = s_1^* \right\}.$$

In each case, calculation of the right-hand-side is facilitated by the asymptotic result that, assuming equal allocation under the null hypothesis,

$$\begin{pmatrix} S_1(t) \\ S_{1,2}(t) - S_1(t) \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_1(t)/4 & 0 \\ 0 & \{D_{1,2}(t) - D_1(t)\}/4 \end{pmatrix} \right), \quad (9)$$

for $t \in [0, T]$, where T is sufficiently large such that all events of interest occur prior to T .

One remaining subtlety is that $E_{H_0} \left\{ \psi_{x_1^{\text{int}}} | S_1(T_{1,2}^*) = s_1 \right\}$ can only be calculated at calendar time $T_{1,2}^*$, where $T_{1,2}^* > T^{\text{int}}$. Determination of b^* must therefore be postponed until this later time.

It is shown in Appendix A that $\psi_{X_1^{\text{int}}} = 1$ if and only if $Z > \Phi^{-1}(1 - \alpha)$, where Z is defined as in (1) with p_1 defined as in (4), T_1 defined as equal to $T_{1,2}$, the second-stage p-value function defined as

$$p_2(Y) = 1 - \Phi \left[2 \left\{ S_{1,2}(T_{1,2}^*) - S_1(T_{1,2}^*) \right\} / \{d_{1,2}^* - D_1(T_{1,2}^*)\}^{1/2} \right], \quad (10)$$

and the specific choice of weights:

$$w_1 = \{D_1(T_{1,2})/d_{1,2}\}^{1/2} \text{ and } w_2 = [\{d_{1,2} - D_1(T_{1,2})\}/d_{1,2}]^{1/2}. \quad (11)$$

Remark 1. In a sense, the Irle and Schäfer method can be thought of as a special case of the Jenkins et al. method, with a clever way of implicitly defining the weights and the end of first-stage follow-up, T_1 . It has two potential advantages. Firstly, the timing of the interim analysis need not be prespecified – in theory, one is permitted to monitor the accumulating data and at any moment decide that design changes are necessary. Secondly, if no changes to the design are necessary, i.e., the trial completes as planned at calendar time $T_{1,2}$, then the original test (6) is performed. In this special case, no data is “thrown away”.

Remark 2. From first glance at (7), it may appear that the data from first-stage patients, accumulating after $T_{1,2}$, is never “thrown away”. However, this data is still effectively ignored. We have shown that the procedure is equivalent to a p-value combination approach where p_1 depends only on data available at time $T_1 := T_{1,2}$. In addition, the distribution of p_2 is asymptotically independent of the data from first-stage patients: note that $S_{1,2}(T_{1,2}^*) - S_1(T_{1,2}^*)$ and $S_2(T_{1,2}^*)$ are asymptotically equivalent (Irle and Schäfer, 2012, remark 1). The procedure therefore fits our description of a “patient-wise separation” design, c.f. Section 2.4, and the picture is the same as in Figure 3. The first-stage patients have in effect been censored at $T_{1,2}$, despite having been followed-up for longer.

This fact has important implications for the choice of $d_{1,2}^*$. If one chooses $d_{1,2}^*$ based on conditional power arguments, one should be aware that the effective sample size has not increased by $d_{1,2}^* - d_{1,2}$. Rather, it has increased by $d_{1,2}^* - d_{1,2} - \{D_1(T_{1,2}^*) - D_1(T_{1,2})\}$, which could be very much smaller.

Remark 3. A potential disadvantage of the Irle and Schäfer (2012) method is that it is not possible to decrease the number of events (nor decrease the recruitment rate) at the interim analysis, as one must observe at least $d_{1,2}$ events (in the manner specified by the original design) to be able to calculate the conditional error probability $E_{H_0} \{\varphi \mid S_1(T_{1,2})\}$.

In addition, one is not permitted to increase the recruitment rate following the interim analysis, nor to prolong the recruitment period beyond that prespecified by the original design. In order to allow such design changes, a small extension is necessary. While the conditional error probability remains $E_{H_0} \{\varphi \mid S_1(T_{1,2})\}$, the second-stage data must be split into two parts, $Y = \{(X_1, X_2) \setminus X_1^{\text{int}}, Y^+\}$, where Y^+ consists of responses from an additional cohort of patients, not specified by the original design (see Figure 4). The second-stage test (7) can be replaced with, e.g.,

$$\psi_{x_1^{\text{int}}}(Y) = \mathbf{1} \left[2S_{2,+}(T_{2,+}) / \{D_{2,+}(T_{2,+})\}^{1/2} \geq b^* \right],$$

where $D_{2,+}(T_{2,+})$ and $S_{2,+}(T_{2,+})$ are the observed number of events and the usual logrank score statistic, respectively, based on the responses of all patients recruited post T^{int} , and $T_{2,+} := \min \{t : D_{2,+}(t) = d_{2,+}\}$ for some $d_{2,+}$ defined at time T^{int} . Again, determination of b^* must be postponed until time $\max(T_{1,2}, T_{2,+})$.

3.3 Effect of unspecified follow-up data

Continuing with the set up and notation of Section 3.1 (which we have shown also fits the Irle and Schäfer (2012) method), the adaptive test statistic is

$$\begin{aligned} Z &= w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2) \\ &= 2w_1 S_1(T_1) / D_1(T_1)^{1/2} + w_2 \Phi^{-1}(1 - p_2). \end{aligned} \tag{12}$$

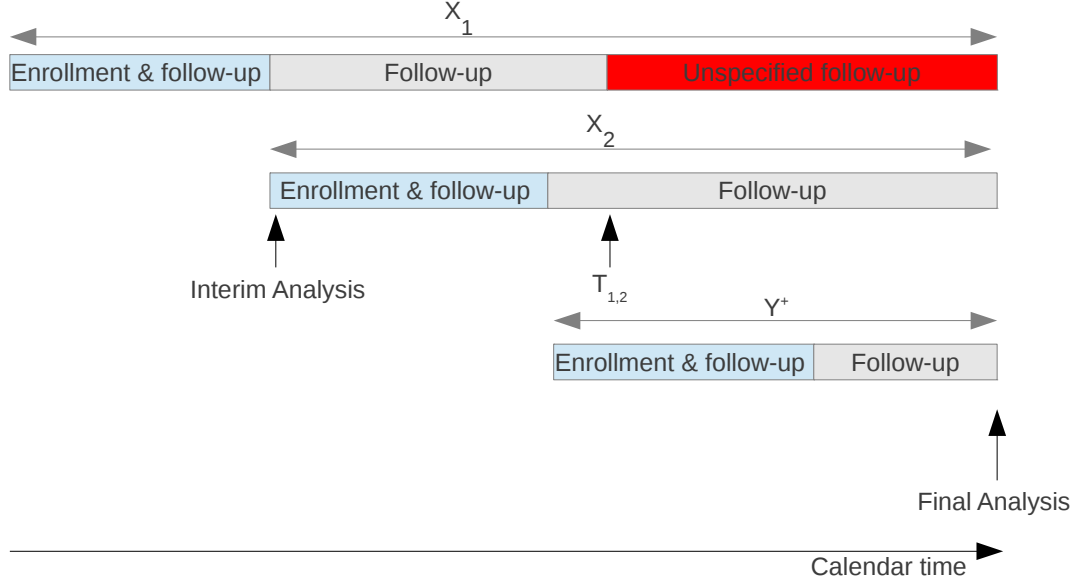


Figure 4: Extension of the Irle and Schäfer approach to allow a prolonged recruitment period.

Suppose, however, that the trial continues until calendar time T^* , where $T^* > T_1$. Strictly speaking, the data from first-stage patients – those patients recruited prior to T^{int} – accumulating between times T_1 and T^* should be “thrown away”. In this section we will investigate what happens, in a worst case scenario, if this illegitimate data is naively incorporated into Z . Specifically, we find the maximum type I error associated with the test statistic

$$Z^* = 2w_1 S_1(T^*)/D_1(T^*)^{1/2} + w_2 \Phi^{-1}(1 - p_2). \quad (13)$$

Since in practice T^* depends on the interim data in a complicated way, the null distribution of (13) is unknown. One can, however, consider properties of the stochastic process

$$Z(t) = 2w_1 S_1(t)/D_1(t)^{1/2} + w_2 \Phi^{-1}(1 - p_2), \quad t \in [T_1, T^{\max}].$$

In other words, we consider continuous monitoring of the logrank statistic based on first-stage patient data. The worst-case scenario assumption is that the responses on short-term secondary endpoints, available at the interim analysis, can be used to predict the exact calendar time the process $Z(t)$ reaches its maximum. In this case, one could attempt to engineer the second stage design such that T^* coincides with this timepoint, and the worst-case type I error rate is therefore

$$P_{H_0} \left\{ \max_{t \geq T_1} Z(t) > \Phi^{-1}(1 - \alpha) \right\}. \quad (14)$$

Although the worst-case scenario assumption is clearly unrealistic, (14) serves as an upper bound on the type I error rate. It can be found approximately via standard Brownian motion results.

Define the *information time* at calendar time t to be $u = D_1(t)/D_1(T^{\max})$, and let $S_1(u)$ denote the logrank score statistic based on first-stage patients, followed-up until information time u . It can be shown that $B(u) := 2S_1(u)/\{D_1(T^{\max})\}^{1/2}$ behaves asymptotically like a Brownian motion with drift $\xi := \theta\{D_1(T^{\max})/4\}^{1/2}$ (Proschan et al., 2006, p. 101).

We wish to calculate

$$P_{\theta=0} \left\{ \max_{t \geq T_1} Z(t) > \Phi^{-1}(1 - \alpha) \right\} = \int_0^1 P_{\theta=0} \left[\max_{u=u_1}^1 B(u) > u^{1/2} w_1^{-1} \{ \Phi^{-1}(1 - \alpha) - w_2 \Phi^{-1}(1 - p_2) \} \right] dp_2, \quad (15)$$

where $u_1 = D_1(T_1)/D_1(T^{\max})$. While the integrand on the right-hand-side is difficult to evaluate exactly, it can be found to any required degree of accuracy by replacing the square root stopping boundary with a piecewise linear boundary (Wang and Pötzelberger, 1997). Some further details are provided in Appendix B.

The two parameters that govern the size of (14) are w_1 and u_1 . Larger values of w_1 reflect an increased weighting of the first-stage data, which increases the potential inflation. In addition, a low value for u_1 increases the window of opportunity for stopping on a random high. Figure 5 shows that for a nominal $\alpha = 0.025$ level test, the worst-case type I error can be up to 15% when $u_1 = 0.1$ and $w_1 = 0.9$. As $u_1 \rightarrow 0$ the worst-case type I error rate tends to 1 for any value of w_1 (see, e.g., Proschan et al., 1992).

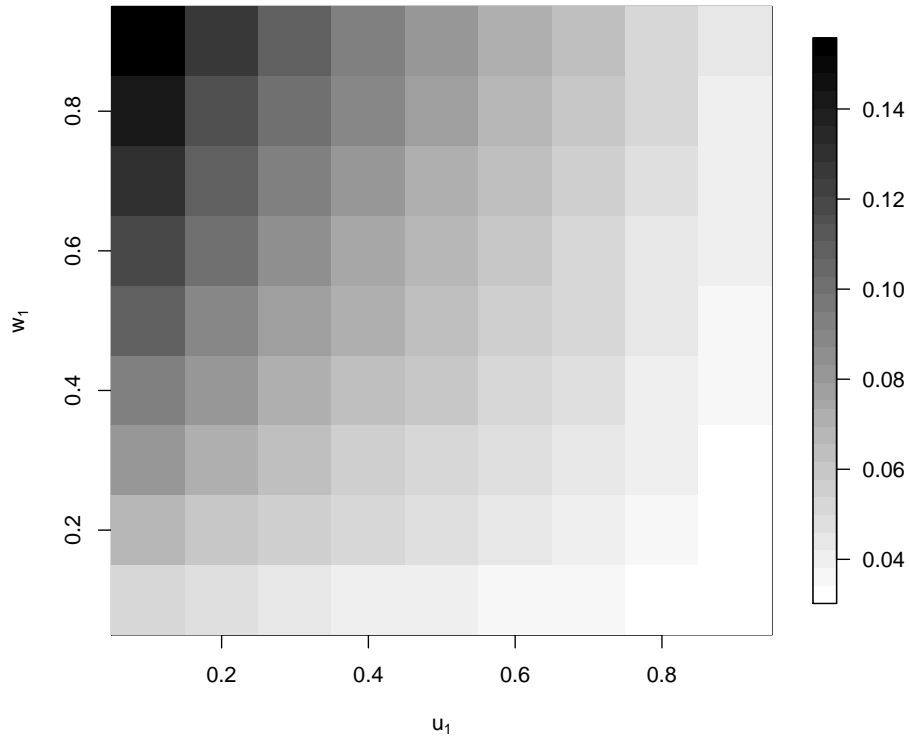


Figure 5: Worst case type I error for various choices of weights and information fractions.

4 Example

The upper bound on the type I error rate, as depicted in Figure 5, varies substantially across w_1 and u_1 . The following example, simplified from Irle and Schäfer (2012), is intended to give an indication of what can be expected in practice.

A randomized trial is set up to compare chemotherapy (C) with a combination of radiotherapy and chemotherapy (E). The anticipated median survival time on C is 14 months. If E were to increase the median survival time to 20 months then this would be considered a clinically relevant improvement. Assuming exponential survival times, this gives anticipated hazard rates $\lambda_C = 0.050$ and $\lambda_E = 0.035$, and a target log hazard ratio of $\theta_R = -\log(\lambda_E/\lambda_C) = 0.36$. If the error rates for testing $H_0 : \theta = 0$ against $H_a : \theta = \theta_R$ are $\alpha = 0.025$ (one-sided) and $\beta = 0.2$, the required number of deaths (assuming equal allocation) is

$$d_{1,2} = 4 [\{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\} / \theta_R]^2 \approx 248.$$

If 8 patients per month are recruited at a uniform rate throughout an initial period of 40 months, and the survival times of these patients are followed-up for an additional 20 months after the end of this period, then standard sample size formulae (Machin et al., 1997, Section 9.2.3.) tell us we can expect to observe around 250 deaths by the time of the final analysis.

Now imagine, as Irle and Schäfer (2012) did, that an interim look is performed after 60 deaths, observed 23 months after the start of the trial. At this point in time, 190 patients have been recruited. Based on the interim results, it is decided to increase the total required number of events from $d_{1,2}$ to $d_{1,2}^*$.

At the time of the 248th death, i.e., the originally planned study end $T_{1,2}$, suppose we observe that 170 of these deaths have come from patients recruited prior to the interim look. We have our weights (11),

$$w_1 = (170/248)^{1/2} \text{ and } w_2 = (78/248)^{1/2}.$$

At this point we make a note of the standardized first-stage logrank score statistic $S_1(T_1) := S_1(T_{1,2})$ and hence p_1 from (4), and continue to follow-up survival times until a total of $d_{1,2}^*$ deaths have been observed. Once these additional deaths have been observed, p_2 can be found from (10), and combined with p_1 to give the adaptive test statistic (1).

Notice that $w_1 = (170/248)^{1/2}$ and, ignoring any potential censoring, $u_1 = D_1(T_1)/D_1(T^{\max}) = 170/190$. In this case a naive application of the test statistic (13) leads to an upper bound on the type I error rate of 0.040. The inflation is not enormous, owing to the relatively slow recruitment rate, but it is not hard to imagine more worrying scenarios.

Suppose, for example, that the trial design called for 48 patients to be recruited per month for 12 months, with 8 months of additional follow-up. Further suppose that an interim analysis took place 6 months into the trial, by which time 288 patients had been recruited, and a decision was made to increase the total number of events. Given the anticipated λ_C and λ_E , a plausible scenario is that 147 of the first 248 events come from first-stage patients, implying that $w_1 = (147/248)^{1/2}$ and $u_1 = 147/288$. This gives an upper bound (14) of 0.066.

4.1 An alternative level- α test

A possible rationale for using (13), instead of (12), is that the final test statistic takes into account all available survival times, i.e., does not ignore any data. If one is unprepared to give up the

guarantee of type I error control, an alternative test can be found by increasing the cut-off value for Z^* from $\Phi^{-1}(1 - \alpha)$ to k^* such that

$$\int_0^1 P_{\theta=0} \left[\max_{u=u_1}^1 B(u) > u^{1/2} w_1^{-1} \{k^* - w_2 \Phi^{-1}(1 - p_2)\} \right] dp_2 = \alpha$$

This will, of course, have a knock on effect on power. Table 1 gives an impression of how much the cutoff is increased from 1.96 when $\alpha = 0.025$ (one sided).

Table 1: Cutoff values for corrected level-0.025 test.

		u_1								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
w_1	0.1	2.29	2.25	2.21	2.19	2.16	2.13	2.11	2.08	2.04
	0.2	2.41	2.35	2.31	2.27	2.23	2.20	2.16	2.12	2.07
	0.3	2.50	2.43	2.38	2.34	2.30	2.25	2.21	2.16	2.10
	0.4	2.58	2.50	2.44	2.39	2.34	2.30	2.25	2.19	2.12
	0.5	2.64	2.56	2.49	2.44	2.38	2.33	2.27	2.21	2.14
	0.6	2.70	2.60	2.53	2.47	2.42	2.36	2.30	2.23	2.15
	0.7	2.74	2.64	2.57	2.51	2.45	2.39	2.33	2.26	2.17
	0.8	2.79	2.68	2.60	2.54	2.48	2.41	2.35	2.28	2.18
	0.9	2.83	2.72	2.64	2.57	2.50	2.43	2.37	2.29	2.19

In assessing the effect on power, at least four probabilities appear relevant:

- A. $P_{\theta=\theta_R} \left[2w_1 S_1(T_1) / \{D_1(T_1)\}^{1/2} + w_2 \Phi^{-1}(1 - p_2) > \Phi^{-1}(1 - \alpha) \right]$.
- B. $P_{\theta=\theta_R} \left[2w_1 S_1(T_1) / \{D_1(T_1)\}^{1/2} + w_2 \Phi^{-1}(1 - p_2) > k^* \right]$.
- C. $P_{\theta=\theta_R} \left[2w_1 S_1(T^{\max}) / \{D_1(T^{\max})\}^{1/2} + w_2 \Phi^{-1}(1 - p_2) > k^* \right]$.
- D. $P_{\theta=\theta_R} \{ \max_{t \geq T_1} Z(t) > k^* \}$.

Power definition **A** corresponds to the “correct” adaptive test. **B** can be thought of as a lower bound on the power of the alternative level- α test, where one conscientiously specifies the increased cutoff value k^* (in anticipation of unpredictable end of first-stage follow-up), but it then turns out that the trial finishes at the prespecified time point anyhow, i.e., $T^* = T_1$. Definition **C** can be thought of as the power of the alternative test if the trial is always prolonged such that all first-stage events are observed. Definition **D**, on the other hand, can be interpreted as the power of the alternative level- α test, taken at face value. In other words, assuming that one takes the opportunity to stop follow-up of first-stage patients when $Z(t)$ is at its maximum. This can be calculated using the same techniques as in Section 3.3.

Figure 6 shows the power of the trial described in Example 4, according to **A-D**. The power has been evaluated conditional on p_2 , as this is a random variable common to all four definitions. The increased cutoff value of the alternative level- α test leads to a sizeable loss of power if the trial completes as planned. In the second scenario at least, the loss of power can be more than made up

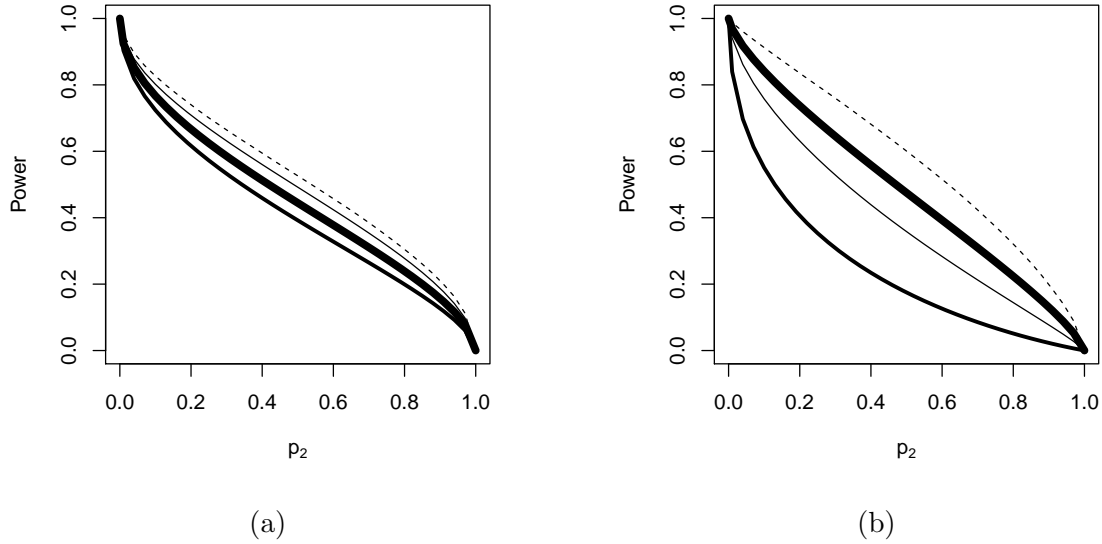


Figure 6: Conditional power as defined by **A** (thin line), **B** (medium line), **C** (thick line) and **D** (dashed line) given p_2 , under the two scenarios described in Example 4. Scenario (a): $D_1(T_1) = 170$, $D_1(T^{\max}) = 190$, $w_1 = (170/248)^{1/2}$ and $\theta_R = 0.36$. Scenario (b): $D_1(T_1) = 147$, $D_1(T^{\max}) = 288$, $w_1 = (147/248)^{1/2}$ and $\theta_R = 0.36$.

for when the trial is prolonged. However, if there is an *a-priori* reasonable probability of prolonging the trial, then one could just start with a larger sample size/ required number of events.

In general, the differences between power definitions **A-D** will tend to follow the same pattern as in Figure 6. The degree to which they differ will depend on w_1 , $D_1(T_1)$, $D_1(T^{\max})$ and θ_R . Intuitively, larger w_1 and smaller u_1 will lead to a greater loss of power going from **A** to **B**, but with a greater potential gain in power going from **B** to **C** (or **D**). The actual gain in power from **B** to **C** (or **D**) will be greatest for large values of θ_R .

4.2 Diverging hazard rates

Consider the second trial design in Section 4, where recruitment proceeds at a uniform rate of 48 patients per month for 12 months, with 8 months of additional follow-up. Suppose, however, that the true hazard rates are not proportional. Rather, $h_E(\tau) = 0.04$ and $h_C^{-1}(\tau) = 0.04^{-1} - 0.6\tau$ for $\tau \in (0, 30)$, where τ denotes the time in calendar months since randomization. Simulating a realization of this trial, 295 patients are recruited in the first six months, by which time there have been 18 deaths on C , and 15 deaths on E . Suppose that at this point it is decided to increase the target number of events from $d_{1,2} = 248$ to $d_{1,2}^* = 350$. At time $T_{1,2}$, the number of deaths from first-stage patients – those patients recruited in the first six months – is $D_1(T_{1,2}) = 151$, such that $w_1 = (151/248)^{1/2}$, $u_1 = 151/295$ and $k^* = 2.41$. The logrank score statistics based on first-stage patients is $S_1(T_{1,2}) = 7.6$, giving a first-stage p-value (4) of

$$p_1 = 1 - \Phi \{2(7.6)/151^{1/2}\} = 0.108$$

and a conditional error probability (6) of $E_{H_0} \{ \varphi \mid S_1(T_{1,2}) = 7.6 \} = 0.213$, using (9). The survival data at time $T_{1,2}^*$, occurring approximately 26 months into the trial, is plotted in Figure 7. On the left-hand-side, all survival times have been included in the Kaplan-Meier curves. There is an obvious divergence in the survival probabilities on the two treatments. However, the test decision of Irle and Schäfer (2012) may only use the data as depicted on the right-hand-side, where the survival times of first-stage patients have been censored at time $T_{1,2}$. They are liable to reach an inappropriate conclusion. In this case, 199 out of the first 350 events are from patients recruited in the first six months, and the logrank score statistic based on first-stage patients is $S_1(T_{1,2}^*) = 16$. The new cutoff value b^* must be found to solve

$$E_{H_0} \left\{ \psi_{x_1^{\text{int}}} \mid S_1(T_{1,2}^*) = 16 \right\} = P_{H_0} \left\{ 2S_{1,2}(T_{1,2}^*)/350^{1/2} \geq b^* \mid S_1(T_{1,2}^*) = 16 \right\} = 0.213,$$

which gives $b^* = 2.76$, using (9). The logrank statistic based on all survival times at $T_{1,2}^*$ is $S_{1,2}(T_{1,2}^*) = 25$ and the test decision is

$$\psi_{x_1^{\text{int}}} = \mathbf{1} \left\{ 2(25)/350^{1/2} \geq 2.76 \right\} = 0,$$

i.e., one cannot reject the null hypothesis. As shown in Section 3.2, the same decision could have been reached by finding the second-stage p-value (10)

$$p_2 = 1 - \Phi \left[2 \{ 25 - 16 \} / (350 - 199)^{1/2} \right] = 0.071,$$

computing the adaptive test statistic (1),

$$Z = w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2) = 1.88,$$

and comparing with $\Phi^{-1}(1 - \alpha) \approx 1.96$. The number of events that have been ignored in making this decision is $D_1(T_{1,2}^*) - D_1(T_{1,2}) = 48$.

If, on the other hand, one had prespecified the alternative test of Section 4.1, then one would be permitted to replace $\Phi^{-1}(1 - p_1)$ in the adaptive test statistic with the value of the standardized logrank statistic at time $T_{1,2}^*$. In this case one would be able to reject the null hypothesis, as

$$Z(T_{1,2}^*) = 2w_1 S_1(T_{1,2}^*)/D_1(T_{1,2}^*)^{1/2} + w_2 \Phi^{-1}(1 - p_2) = 2.69 > k^*.$$

5 Discussion

Adaptive design methodology – developed over the past two decades to cope with mid-study protocol changes in confirmatory clinical trials – is becoming increasingly accepted by regulatory agencies (Elsaesser et al., 2013). It is therefore unfortunate that the important case of time-to-event data is not easily handled by the standard theory. As far as survival data are concerned, all proposed solutions have limitations and there is a trade-off between strict type I error control, power, flexibility, the use of all interim data to substantiate interim decision making, and the use of all available data in making the test decision at the final analysis.

The proposed solutions of Irle and Schäfer (2012) and Jenkins et al. (2011) offer strict type I error control and allow full use of the interim data. The Jenkins et al. (2011) method allows one to

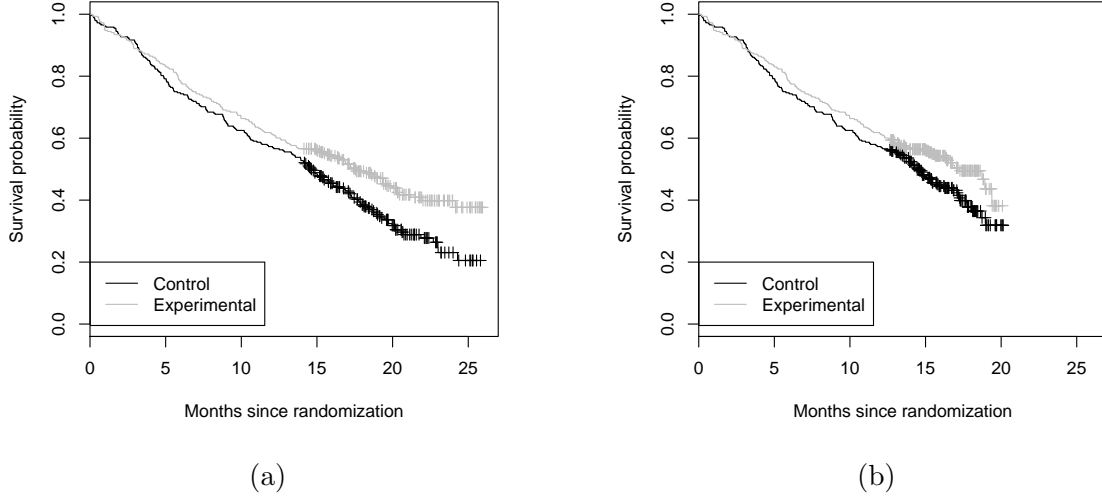


Figure 7: Kaplan-Meier plots corresponding to the hypothetical trial described in Section 4.2 based on (a) all available survival times at $T_{1,2}^*$, and (b) with first-stage and second-stage patients censored at $T_{1,2}$ and $T_{1,2}^*$, respectively.

change the recruitment rate at the interim analysis – something that is disallowed by Irle and Schäfer (2012), where one is only permitted to increase the observation time. On the other hand, Irle and Schäfer (2012) is more flexible in the sense that the timing of the interim analysis need not be prespecified. In both cases, the final test decision only depends on a subset of the recorded survival times, i.e., part of the observed data is ignored. This is usually deemed unacceptable by regulators. Furthermore, it is the long-term data of patients recruited prior to the interim analysis that is ignored, such that more emphasis is put on early events in the final decision making. This neglect becomes egregious when there is specific interest in learning about the long-term parts of the survival curves.

We have therefore proposed an alternative procedure which offers the same type I error control and flexibility as Jenkins et al. (2011) and Irle and Schäfer (2012), in addition to a final test statistic that takes into account all available survival times. However, in order to achieve this, a worst-case adjustment is made a-priori in the planning phase. If no design modifications are performed at the interim analysis, the worst-case critical boundary must nevertheless be applied. This results in a loss of power.

Methods based on the independent increments assumption have been only briefly mentioned in Section 2.3. They suffer from the limitation that decision makers must be blinded to short-term data at the interim analysis. On the other hand, subject to this blinding being imposed, the type I error rate is controlled and the final test decision is based on all available survival times. This could therefore be a viable option in situations where the short-term data is sparse or relatively uninformative. Yet another option, if one is prepared to give up strict type I error control, is simply to use the usual logrank test at the final analysis. The true operating characteristics of such a procedure are unclear, owing to the complex dependence on the interim data.

Our alternative level- α test may have practical applications in multi-arm survival trials (Jaki and

Magirr, 2013) and adaptive enrichment designs. In this case, one must take great care in applying the methodology of Jenkins et al. (2011) or Irle and Schäfer (2012). One specific issue is that dropping a treatment arm will affect recruitment rates on other arms. Also, for treatment regimes that have to be given continuously over a period of time, it would be unethical to keep treating patients on treatment arms that have been dropped for futility. This may affect the timing of analyses on other arms. Incorporating some flexibility into the end of patient follow-up could confer advantages here. More research is needed in this area.

The usefulness of performing design modifications has to be thoroughly assessed on a case-by-case basis in the planning phase. Interim data may be highly variable, and the interim survival results may be driven mainly by early events. Consequently, the interim data may be too premature to allow a sensible interpretation of the whole survival curves and may not be a reliable basis for adaptations.

In this respect, the best advice might be to thoroughly assess the characteristics of adaptive trial designs in comparison with more standard approaches, and to plan for adaptations only in settings where the advantages are compelling. If in the planning phase there is a strong likelihood that the number of patients will need to be increased, or the observation time extended, our analysis has shown that there is no uniformly best design. All proposals to implement adaptive survival designs have their limitations. If the main objective is strict type I error control when using all data, then our proposal should be considered as a valid option.

Appendix A

Connection between conditional error and combination test

The cut-off b^* satisfies

$$\begin{aligned} E_{H_0} \{ \varphi \mid S_1(T_{1,2}) = s_1 \} &= P_{H_0} \{ 2S_{1,2}(T_{1,2}^*) / (d_{1,2}^*)^{1/2} \geq b^* \mid S_1(T_{1,2}^*) = s_1^* \} \\ &= P_{H_0} \left[2 \{ S_{1,2}(T_{1,2}^*) - S_1(T_{1,2}^*) \} / \{ d_{1,2}^* - D_1(T_{1,2}^*) \}^{1/2} \geq c^* \mid S_1(T_{1,2}^*) = s_1^* \right], \end{aligned}$$

which implies that $c^* = \Phi^{-1} [1 - E_{H_0} \{ \varphi \mid S_1(T_{1,2}) = s_1 \}]$, using (9). Therefore,

$$\begin{aligned} \psi_{X_1^{\text{int}}} = 1 &\Leftrightarrow 2S_{1,2}(T_{1,2}^*) / (d_{1,2}^*)^{1/2} \geq b^* \\ &\Leftrightarrow 2 \{ S_{1,2}(T_{1,2}^*) - S_1(T_{1,2}^*) \} / \{ d_{1,2}^* - D_1(T_{1,2}^*) \}^{1/2} \geq c^* \\ &\Leftrightarrow \Phi^{-1}(1 - p_2) \geq \Phi^{-1} [1 - E_{H_0} \{ \varphi \mid S_1(T_{1,2}) = s_1 \}] \\ &\Leftrightarrow p_2 \leq E_{H_0} \{ \varphi \mid S_1(T_{1,2}) = s_1 \}. \end{aligned}$$

The conditional error probability, $E_{H_0} \{ \varphi \mid S_1(T_{1,2}) = s_1 \}$, can be found from the joint distribution (9) at calendar time $T_{1,2}$. Omitting the argument $T_{1,2}$ from S_1 , $S_{1,2}$, D_1 and $D_{1,2}$:

$$\begin{aligned} E_{H_0} \{ \varphi \mid S_1 = s_1 \} &= P_{H_0} \left\{ 2S_{1,2} / (D_{1,2})^{1/2} > \Phi^{-1}(1 - \alpha) \mid S_1 = s_1 \right\} \\ &= P_{H_0} \left[2(S_{1,2} - S_1) / (D_{1,2} - D_1)^{1/2} > \Phi^{-1}(1 - \alpha) \{ D_{1,2} / (D_{1,2} - D_1) \}^{1/2} \right. \\ &\quad \left. - 2S_1 / (D_{1,2} - D_1)^{1/2} \mid S_1 = s_1 \right] \\ &= 1 - \Phi \left[\Phi^{-1}(1 - \alpha) \{ D_{1,2} / (D_{1,2} - D_1) \}^{1/2} - \Phi^{-1}(1 - p_1) \{ D_1 / (D_{1,2} - D_1) \}^{1/2} \right] \end{aligned}$$

and therefore $p_2 \leq E_{H_0} \{\varphi \mid S_1(T_{1,2}) = s_1\}$ if and only if

$$\{D_1(T_{1,2})/d_{1,2}\}^{1/2} \Phi^{-1}(1 - p_1) + [\{d_{1,2} - D_1(T_{1,2})\}/d_{1,2}]^{1/2} \Phi^{-1}(1 - p_2) \geq \Phi^{-1}(1 - \alpha).$$

Appendix B

Computation of (14)

For simplicity, consider replacing the square root boundary in (15) with a linear boundary. Conditional on p_2 , our problem is to find $P_{\theta=0} \{B(u) < au + b, u_1 < u \leq 1\}$, where a and b are found by drawing a line through

$$u_1, u_1^{1/2} w_1^{-1} \{\Phi^{-1}(1 - \alpha) - w_2 \Phi^{-1}(1 - p_2)\}$$

and

$$1, w_1^{-1} \{\Phi^{-1}(1 - \alpha) - w_2 \Phi^{-1}(1 - p_2)\}.$$

For constants a, b and c , with $b, c > 0$, Siegmund (1986) shows that

$$P_{\theta=0} \{B(u) \geq au + b, \text{ for some } 0 < u \leq c \mid W(c) = x\} = \exp \{-2b(ac + b - x)/c\}$$

and integrating over x gives

$$P_{\theta=0} \{B(u) < au + b, u \leq c\} = \Phi \{(ac + b)/c^{1/2}\} - \exp(-2ab) \Phi \{(ac - b)/c^{1/2}\}. \quad (16)$$

Therefore, conditioning on the value of $B(u_1)$,

$$\begin{aligned} P_{\theta=0} \{B(u) < au + b, u_1 < u \leq 1\} &= \int_{-\infty}^{au_1} P_{\theta=0} \{B(u) < au + b, u_1 < u \leq 1 \mid B(u_1) = x\} dP_{u_1}(x; 0) \\ &= \int_{-\infty}^{au_1} P_{\theta=0} \{B(u + u_1) - x \\ &\quad < a(u + u_1) + b - x, 0 < u \leq 1 - u_1 \mid B(u_1) = x\} dP_{u_1}(x; 0) \\ &= \int_{-\infty}^{au_1} P_{\theta=0} \{B(v) < av + au_1 + b - x, 0 < v \leq 1 - u_1\} dP_{u_1}(x; 0) \\ &= \int_{-\infty}^{au_1} \Phi \left\{ \frac{a + b - x}{(1 - u_1)^{1/2}} \right\} \\ &\quad - \exp \{-2a(au_1 + b - x)\} \Phi \left\{ \frac{a(1 - 2u_1) - b + x}{(1 - u_1)^{1/2}} \right\} dP_{u_1}(x; 0). \end{aligned}$$

Greater accuracy can be achieved by replacing the square root boundary with a piece-wise linear boundary, in which case one must condition on the value of the Brownian motion at each of the cut-points (Wang and Pötzelberger, 1997).

References

- Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50:1029–1041. Correction: *Biometrics* 1996; 52:380.
- Bauer, P. and Posch, M. (2004). Letter to the editor. *Statistics in Medicine*, 23:1333–1334.
- Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.
- Brannath, W., Gtjhr, G., and Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association*, 107:824–832.
- Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in medicine*, 28(10):1445–1463.
- Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28:1181–1217.
- Cox, D. R. and Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.
- Desseaux, K. and Porcher, R. (2007). Flexible two-stage design with sample size reassessment for survival trials. *Statistics in medicine*, 26(27):5002–5013.
- Di Scala, L. and Glimm, E. (2011). Time-to-event analysis with treatment arm selection at interim. *Statistics in medicine*, 30(26):3067–3081.
- Elsaesser, A., Regnstroem, J., Vetter, T., Koenig, F., Hemmings, R., Greco, M., Papaluca-Amati, M., and Posch, M. (2013). Adaptive designs in european marketing authorisation a survey of advice letters at the european medicines agency. *Submitted*.
- Friede, T., Parsons, N., and Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*, 31(30):4309–4320.
- Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J., and Nicholas, R. (2011). Designing a seamless phase ii/iii clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in medicine*, 30(13):1528–1540.
- Hampson, L. V. and Jennison, C. (2013). Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):3–54.
- Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal*, 43(5):581–589.
- Irle, S. and Schäfer, H. (2012). Interim design modifications in time-to-event studies. *Journal of the American Statistical Association*, 107:341–348.

- Jahn-Eimermacher, A. and Ingel, K. (2009). Adaptive trial design: A general methodology for censored time to event data. *Contemporary clinical trials*, 30(2):171–177.
- Jaki, T. and Magirr, D. (2013). Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. *Statistics in Medicine*, 32(7).
- Jenkins, M., Stone, A., and Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10:347–356.
- Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall.
- Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):pp. 1286–1290.
- Liu, Q. and Pledger, G. W. (2005). Phase 2 and 3 combination designs to accelerate drug development. *Journal of the American Statistical Association*, 100(470):493–502.
- Machin, D., Campbell, M., Frayers, P., and Pinol, A. (1997). Sample size tables for clinical trials. *Blackwell Science, Cambridge*.
- Müller, H. H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57:886–891.
- Posch, M. and Bauer, P. (1999). Adaptive Two Stage Designs and the Conditional Error Function. *Biometrical Journal*, 41:689–696.
- Proschan, M. A., Follmann, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48:1131–1143.
- Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51:1315–1324.
- Proschan, M. A., Lan, K. K. G., and Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials*. New York: Springer.
- Schäfer, H. and Müller, H.-H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in medicine*, 20(24):3741–3751.
- Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *Annals of Statistics*, 14:361–404.
- Stallard, N. (2010). A confirmatory seamless phase ii/iii clinical trial design incorporating short-term endpoint information. *Statistics in medicine*, 29(9):959–971.
- Wang, L. and Pötzelberger, K. (1997). Boundary crossing probability for Brownian motion and general boundaries. *Journal of Applied Probability*, 34:54–65.

Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal*, 48(4):714–729.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Chichester: Wiley.